DOCUMENT RESUME

ED 046 247 EM 008 671

AUTHOR Hays, Daniel G.
TITLE Computer Analysis of Descriptions of Classroom Events.
INSTITUTION Missouri Univ., Columbia. Center for Research in Social Behavior.
PUB DATE Feb 71
NOTE 17p.; Paper presented at Annual Meeting of the American Educational Research Association (New York, N.Y., February 1971)

EDRS PRICE EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS Automation, Computer Oriented Programs, Educational Research, *Electronic Data Processing, Information Retrieval, *Language Research, *Social Behavior, Verbal Communication
IDENTIFIERS *Activity Code and Text System, ACTS, Keywords in Context, KWIC

ABSTRACT

A computer-based system for handling transcripts describing social behavior, verbal and otherwise, is described in this paper. Among the features of note in the system are the provision for fairly flexible conventions for noting, in transcripts, several kinds of useful data including actor designations, sequencing indicators, and user-defined systematic annotations. (Author)

Computer Analysis of Descriptions of Classroom Events

Daniel G. Hays

University of Missouri--Columbia

This paper describes some computational aids for handling transcriptions of verbal and other social interaction, and discusses the advantages for analysis which such automation allows. These aids are part of a computer-based system called ACTS, for Activity Code and Text System, which is being developed for the preparation, storage, retrieval, and, eventually, some analysis of transcriptions of human behavior, verbal and otherwise. The system was conceived in the course of work on the analysis of language behavior in the classroom, and it should be of interest to specialists in educational research, since in this field, perhaps more than in any other area in the social sciences, the description and analysis of segments of human behavior as it occurs naturally in real situations has attracted considerable interest.

If transcripts or other systematic descriptions of behavior are prepared and analyzed in other than a cursory or impressionistic way; and particularly if the amount of data of this sort is at all large, the facility of a suitably programmed computer for keeping track of data items can be very helpful. Finding all occurrences of a given word, for instance, in even a twenty page transcript can be very tedious for a human being, and such chores are likely to lead not only to disgruntlement but also to errors. A computer, on the other hand, can perform such bookkeeping chores very quickly, and free the human investigators for activities for which they are more suited, such as making semantic

judgements or discovering patterns which depend on cues so subtle that computers as we presently understand how to program them are not much help. Moreover, once the data are stored in the computer, other kinds of analysis possible with computers such as some kinds of content analysis and the examination of some sequential patterns in the data--can be performed on them which might be as useful to the researcher as the initial "simple" information retrieval.

In a number of fields in the social sciences, investigators are interested in behavior in situ, as it occurs naturally, as contrasted with laboratory behavior from which only a few items of data are abstracted in the context of some experimental design. It is in the field of educational research where perhaps the most extensive investigation of situated behavior has occurred, however. The number of observational studies of classroom behavior is quite large, and a good deal of thought has gone into the description of classroom events, making judgements about these events, and seeking patterns of these events and relationships among them and other variables. The work of Bellack (1960), Flanders (1962), L. Smith (1968), B. Smith (1967), and their colleagues is well known. The first edition of Mirrors for Behavior, instruments the compendium of work on classroom and related interaction by A. Simon and G. Boyer was very lengthy; and, perhaps the fact that the amount of work reported in the second edition (1970) is approximately double that reported in the first edition (1968), may indicate that investigations in this area continue to be challenging to researchers.

ACTS was designed specially to facilitate such research, where transcriptions or similar descriptions of behavior are prepared or might be prepared, to be consulted in the course of making judgements, or

analyzed in and of themselves. If the analysis of descriptions of the stream of ongoing behavior in some situation of interest is to be at all fine grained, and if any sizeable amount of interaction is analyzed, then the assistance of the computer in this analysis can effect a great savings in human effort. Indeed, some kinds of analysis may be done when the computer is involved which would be either too complicated or too time-consuming otherwise.

In its present stage of development, ACTS allows storage of transcript and related data in such a form that the research can retrieve this information flexibly, either for perusal and comparison, or else for further automatic processing. It consists of (1) routines for input of transcripts, either on punch-cards or using typewriter-like entry via tape cartridges prepared on an IBM MTST; (2) routines for correcting and otherwise adjusting the transcripts once they have been entered; (3) programmed procedures for segmenting the text and labelling these segments according to type of data (word, sentence, actor block, etc.); (4) procedures for setting up directories, or maps, which allow the researcher to get to various parts of his data easily, and which can allow linkages between his basic data and other data, such as files of coded judgements; (5) basic retrieval and output procedures. It is programmed in PL/I and run on an IBM 360/65 with heavy reliance on disc storage.

Basic Data Structuring

When an investigator wishes to use the facilities of ACTS, he must first prepare his transcripts and related data for entry into the computer, taking care to be consistent in such matters as spelling, punctuation, and other conventions which may be relevant to machine processing. From these carefully prepared texts, special representations

of the transcripts are generated in the computer, which are segmented and marked internally in such a way that access to various parts of the data can be had in a flexible way. These representations, called basic text files, are at the heart of ACTS: the fact that the basic data are structured and marked according to the type of the individual pieces of data making up the transcript not only facilitates retrieval, but also allows them to enter investigator-user-supplied analysis routines in a systematic way.

This structuring of the data, which is prerequisite for the setting up of directories and other linkage facilities, is performed by a central program which uses cues present in the transcript as entered into the computer (such as word boundaries, user-defined special brackets, terminating punctuation marks, and so on), together with certain specifications supplied by the investigator regarding optional use of predetermined configurations of symbols. In performing this structuring, several logically separable tasks are performed: (1) the stream of characters, or written symbols, which make up the transcript, is partitioned into stretches of characters which will become the contents of first-order or elementary data-types; (2) these segments are labelled according to the kind of data they represent (e.g., spoken word, punctuation mark, special information segment); and (3) these first-order labelled segments enter into higher-order structures, which indicate the arrangement of the elementary data-type occurrences (such as spoken sentence representation, special annotation with tag and contents, etc.). As will be seen below, some of these higher-order structures are fairly complex.

Thus, what initially was just a long string of symbols entering the computer one after another becomes a string of structures of segments of symbols, where the parts of which are marked for future reference. Without some kind of structuring, the transcript would remain just a string of individual symbols, and any computation involving this string would involve at least partial restructuring. The same is true, of course, for any kind of data entering the computer.

Most researchers, at this stage of our affluence and technology, are familiar with the preparation of numerical or qualitative code data for machine processing. Ordinarily, items of data are punched into specified positions on IBM cards, with strong constraints on what kind of data must go exactly where. (The code letter for sex of respondent, for instance, must be punched in column eleven and no where else; otherwise it will be lost, or a statistical program may abort trying to compute the product of the letter M and a test score.) For data prepared in the above way, the segmentation of the symbols on the punchcard, and their identification as to type, is performed by the programs which process them according to formatting specifications which depend on exact location and length of the strings of symbols.

Textual data, on the other hand, cannot conveniently be constrained to such fixed format specifications since, by its very nature, the length and location of its units are variable. Furthermore, the identification of type of data may depend on fairly complex contextual cues. ACTS provides for basic segmentation and identification of free-format textual data in a way that allows the user some latitude in choice of punctuation cues, special uses of certain symbols, and so on. If he

wishes to use the symbol string "//", for instance, to indicate boundaries in transcripts of spoken behavior (Garvey, 1970), he may do so: there is no reason that he must be constrained to use ".", "?", or "!". ACTS goes beyond ordinary text-handling systems also in that it is tailored for transcript data, with its peculiarities, rather than being based on the simpler graphical system of straight literary text.

Basic kinds of data. The basic types of data described in this paper, and cues for recognizing them, are essentially those outlined in Hays (1970). Though they are only a subset of the kinds of structures which can be recognized, they represent a fairly manageable package designed to handle the structures likely to be useful in working with transcripts of classroom behavior and similar situations.

The main kinds of data are as follows:

1. Actor designations. In describing social behavior, who is doing the behaving usually must be indicated. An actor designation, in ACTS, in an ACTS transcript, is delimited by user-defined special characters, and may contain one or more actor identifications. If more than one actor is cited as performing some action, let us say in performing a duet, actor names are separated by commas or the word "and".

(Joe) He hit me.

(Henry and Jack) Ya ya ya ya ya. /to the tune of the familiar childhood taunt/

2. Annotations. Comments, systematically constructed or not, are permitted in the transcript, and are set off from the rest of the text by special delimiters or brackets. Example:

(Mary) /slowly/ The capital of Alabama is...

/Mary sticks pencil in her ear./

3. Basic text units. In the scheme described here, whatever is not annotation or actor designation is taken as basic description data. For most applications, it is expected that this data will be some representation of utterances, though it might be interpreted as overt behavioral descriptions in some special language.

a. Basic text words. Any string bounded by blanks or defined punctuation, which does not lie within annotation or actor designation delimiters, is marked as a functional unit.

b. Terminating punctuation. Symbols, or strings of symbols, may be defined by the investigator as constituting two kinds of punctuation units, terminating and non-terminating. Both are recognized in about the same way, relying on their pattern and on some contextual conditions.

A string of basic text units ending with a terminating punctuation, with possibly a quotation character following it, make up the higher-order data-type sentence. Annotations may occur within sentences, or between them, but may not occur within a first-order unit, such as a basic text word. The data-type "sentence" may receive various interpretations. It may be taken as a bounded sequence of behavioral descriptors in a code language, for instance. Or it may refer to a major segmenting of a stretch of spoken language in the usual sense. There is of course not any necessity that the "sentence" be complete in the grammar-book sense.

An actor designation followed by anything, up the next actor designation, makes up an actor block. Normally, an actor block will contain one or more sentences, and may contain annotations.

Sentence discontinuities. Sentences and actor blocks are the main higher-order structures in transcripts. From one point of view, the description of a social situation is a sequence of actor blocks, and the

main interest lies in the flow of activity from one actor to another. In education research, when one is interested primarily in the interactive aspects of the events, one is likely to focus on patterns of actor blocks, and to approach finer analyses involving sentences and words within sentences, in this context. When one is primarily interested in language behavior, however, many analyses will be based on sentences.

If it were the case that sentence boundaries always coincided with actor block boundaries, retrieval and analysis would be simpler than it is. However, people interrupt one another, and sentences are sometimes finished after someone else has said something. (In describing simultaneous events, sentence overlap of actor blocks is sometimes convenient, even though there is no interruption in the usual sense.) For this reason, sentence and actor block structure, and retrieval, are not strictly hierarchical in ACTS.

Annotations. The treatment of annotations in ACTS constitutes one of the main areas of the system's flexibility in meeting the needs of investigators.. Working with transcripts describing the events in naturally occurring situations (in fact, in public school classrooms), it has turned out to be very convenient to include annotations in the text, either for incidental information which will not receive formal analysis but which helps to make the descriptions more understandable, or for information that may be subject to more systematic treatment later, but that is not strictly speaking a part of what was referred to above as "basic text data", such as transliterations of speech. If the transcript consists only of words which are spoken, together with actor identification, it is the case, often enough to be annoying, that this bare record just doesn't make much

sense without auxiliary information. It is often interesting to include auxiliary data in the transcript, in sequence; for instance, in analyzing utterances, one may wish to sort one's data according to the apparent 'target' of the communication, as well as the source or speaker; or a notation concerning the tone of voice of a teacher utterance may be interesting in relation to the verbal content of the subsequent student utterances.

To accommodate such needs, ACTS allows both unsystematic annotations, which are included only for purposes of clarification (or setting down hunches or strictly incidental observations into the record), and systematic annotations (which we will sometimes call keyworded annotations) which are prepared systematically and may be systematically retrieved as well.

An unsystematic annotation is simply a string bounded by annotation brackets. Systematic annotations are bounded by explicit brackets, but have as well a tag or keyword, which labels the annotation according to type. The choice of kayword, and the corresponding creation of an annotation type, is left to the investigator. He has, in other words, the ability to define a number of kinds of data, of relevance to his particular research problems and data characteristics.

For example, in Figure 1, annotations with the kayword "TO" indicate the apparent targets of utterances. Annotations preceded by the string "BD" indicate descriptions of overt behavior, and their contents may be accessed in context for further processing. Having keywords in upper-case letters facilitates their identification when human beings are examining transcripts or parts of them (and also helps prevent errors in the unwitting use of a keyword in what is meant to be an unsystematic annotation),

(T) Yes, Big Man, oh, /lengthened/ he's ferocious. He's fierce. 1

(Lavorah) Wow. 2

(T) He nearly, he nearly tears that cage down. +I like to stand way back from Big Man, though. Alright, [Nancy], what is your favorite animal? 3 4

(Nancy) The horse. /Nancy seems indifferent./ 5

(T) /TO John B/ You are right. /TO class/ Okay, what is this? 6

(Part of Class) /SIM AA/ /SYNC/ A frog. 7

(Part of Class) /SIM AA/ UNIS/ A tiger. 8

(T) /TO Mary B/ What is it, <Mary: S19>? 9

(Mary B) A turtle. 10

(T) It's a turtle /lengthened/. Now say all together, /SIM DD/ turtle. /ESM DD/ 11

(Class) /SIM DD/ /UNIS/ Turtle. /ESM DD/ 12

(T) Alright. /BD T walks to back of room./ Alright, which one is...# 13

/BD T pulls down map of India./ Here it is...# /TO male student 14

near back of classroom/ What country is 15

(---) /INT/ ---. /apparently about an extraneous issue/ 16

(T) this? Uh... [Lavorah]?

Figure 1

Parts of a transcript of an elementary school lesson.

but is not necessary. The annotation "/lengthened/" is actually systematic, since "lengthened" is one of a set of keywords specified for ACTS for this particular set of transcripts, which taken together are used to systematically indicate prosodic characteristics of the spoken language.

An investigator may be interested in pauses, as well as some class of gestural behaviors, in the context of spoken language analysis. For his transcripts, he could define annotation keywords representing kinds of pauses, and a type of systematic annotation for gestural indicators, which would contain restricted content items.

(Teacher) We uh /P/ want to

(Teacher) We want to uh /HES/ make our lessons neatly /P/

Temporal overlap. One of the characteristics of almost any social behavior is that the events tend to overlap in time; and it may be of interest in an investigation to indicate this overlap in the descriptions of the behavior. In the classroom, for instance, a student may start giving an answer before the teacher has finished the question; or the entire class may respond in unison with the teacher. Even in fairly orderly classrooms, more than one student may be speaking at once, in an animated discussion.

ACTS reserves two special annotations to 'bracket' stretches of text representing simultaneous events. Each consists of annotation delimiters such as slashes, a keyword indicating the beginning or ending of a segment which is simultaneous with some other segment, and a two-character string which serves to distinguish each case of overlap. For example, if SIM indicates the start of an overlap, and ESM indicates the termination of an overlap, we might have:

(T) Now this is a /SIM xy/ pork chop. /ESM xy/

(Jim) /SIM xy/ pork chop. /ESM xy/

Simultaneity annotations, which allow retrieval of all examples of overlap, together with the possibility that sentences may span more than one actor block, together with other annotations which may be defined, allow the depiction of the temporal sequencing of events in a not strictly sequential fashion. Since the temporal sequencing and overlap of behavior is a basic part of interaction data, ACTS thus provides facilities for representing in computer structures this aspect of the basic structure of the described events. In work that we have done, the simultaneity annotations, two annotations for unison and asynchronous group responses, an annotation to indicate an interruption, and a special terminating punctuation unit which indicates sentence "left hanging", allow representation of what seem to be the major phenomena of this sort.

Getting to the Data

From the point of view of the user, there are two main ways of accessing some part of the computer representation of a transcript: by location, and by content.

The first, access by location, is fairly straightforward. Working from a printout of a transcript, with actor blocks and sentences numbered, the investigator specifies entry into the data representation at actor block number 54, for instance, or the fifth word in sentence 528. Though it is simple, this sort of access may be useful when transcripts are being examined carefully, scrutinized carefully, and parts of them extracted for further processing using criteria which are subjective. This kind of entry is useful as well when setting up directories which reflect some judged 'episodic bracketting' of the text (Barker and Wright, 1956).

In accessing parts of the text by content, the real advantages of computer retrieval become manifest. Currently, several kinds of content may be used for retrieval purposes in ACTS. A transcript may be entered by (1) actor designation, (2) basic text word, or (3) type of annotation and (4) optionally content of annotation. For instance, one may wish to access all utterances belonging to male students. Or one may wish to find all sentences containing a personal pronoun.

Once a basic text file is entered, by either of the above methods, examination can proceed backwards or forwards in the file, in what might be called locally sequential processing. Or, one may wish to extract a given content item in its immediate context, and go on to the next occurrence of that content item.

Access by location is made possible in an efficient way by index files which give the location in computer storage of each major higher-order unit, in sequence. Access by content is made possible by what are called inverted index files, which for example, contain a list of each distinct basic text word, together with a list of locations in the text file of that word.

Immediate analysis aids. One consequence of the existence of inverted index files is that frequency counts of text words, annotations, and the occurrence of actor blocks associated with particular actors, etc., are available, and do not have to be computed separately. A simple listing of words occurring in a class hour, with their frequency of occurrence, may be of some interest. If much use is made of systematic annotations, frequency statistics on their occurrence may constitute substantive results. Thus, some "analysis" is provided automatically by ACTS.

Another consequence is that the investigator has a flexible tool for selective perusal of his data. In trying to understand classroom interaction (or interaction on the playground or in the home), simply being able to examine, for instance, parts of the data with similar surface content may lead to insights. The value of Key Words in Context (KWIC) arrangement of textual data is well known; ACTS provides selective contextual examination with the ability to specify conditions on what is printed out (for instance, one may be interested only in teacher utterances containing student names; or passages which contain marked interruptions).

Linkages. Index files, or maps of the data, allow the association of files of coded judgements with various parts of the data, at various levels. For instance, one may be interested in coding attributes of actor blocks for one kind of analysis; and coding attributes of sentences or words for other purposes. Most of the work summarized in Simon and Boyer (1970) involves making judgements about events at some level of molarity. If an investigator is working with transcripts, it would appear to be useful to storage judgements about actor blocks and santences, and perhaps words as well. Certainly in any sophisticated analysis system, surface text alone is not enough, given our present understanding of both interaction and language, and the ability to link strings of attributes and their values to basic descriptions should be useful.

Varieties of Transcripts

In the above, we have used examples, and based our discussion, primarily on annotated transcriptions of utterances. Other kinds of texts may be handled by ACTS as well.

For instance, one may work with what might be called a "compressed" transcript, which relies heavily on paraphrase. Extensive notes taken during the observation of a classroom, marked for date, class, teacher, etc., may be stored as transcripts; and access by annotation-type or by word may be useful for later consultation of the notes, especially if the amount of them is sizeable. For instance, retrieval by student name may be interesting.

Another kind of transcript has as its basic text not transliterations of utterances, but behavior descriptions. These may be in some behavioral description language, with a structure and semantics less complicated than that of ordinary English, or even consist of a string of codes. Formally, a string of code symbols in sequence, marked as to actor, and bounded by "punctuation", may have the same characteristics as a string of words reflecting verbal behavior.

The existence of an index file system allows not only coded attributes to be linked to basic text files, but also allows other text. Soskins and Johns protocols (1963), for instance, separate behavioral descriptions and utterances. One might similarly separate descriptions and interpretations added after the fact, or sentences and canonical paraphrases, for instance.

Current Status and Prospects

ACTS may still be described as a system in the process of development, though its basic facilities are in use. Data structuring facilities of somewhat more complexity than those described here are being worked with, and are convenient in some applications involving fine-grained language data. These more elaborate facilities are also somewhat more complicated to use, and expensive in terms of computation.

Another area of development is facilities for flexible retrieval. Basic access mechanisms exist, but since no user-oriented retrieval language has been implemented, they are probably lacking somewhat from the point of view of the investigator who has only a nodding acquaintance with computing.

In developing any data-handling system, needs of the people who will use the system are very important. The kinds of data they will enter into the system, and how they want to get at the data and analyze it, have strong implications for a number of technical aspects of the system as it is developed. It is tempting to assume that one knows what other people want, or need, but often it is not the case that one in fact does.

For this reason, we are most interested in learning more about the kinds of data, and kinds of demands which might be placed on data, from persons who are working with transcriptions or other descriptions of social interaction, verbal or otherwise.

Summary

A computer-based system, ACTS, for handling transcript data in studies of social interaction, has been described, and some uses of the system for retrieving data and structuring it for further analysis have been discussed.

References

Barker, R. G., (1963). The Stream of Behavior: Explorations of its Structure and Content. New York: Appleton-Century-Crofts.

Barker, R. and Wright, H., (1956). Midwest and its Children. Evanston, Ill.: Row, Peterson.

Bellack, A., et al, (1960). The Language of the Classroom. New York: Teachers' College Press, Columbia University.

Flanders, N. A., (1962), "Using Interaction Analysis in the In-service Training of Teachers", Journal of Experimental Education, 30, 313-316.

Garvey, Catherine (1970). The structure of a conversation type. Paper read at Linguistic Society of America Annual Meeting, Washington, D. C., December, 1970.

Hays, D. G. Accessing information in behavioral description analysis. In Berton, Alberta (Ed.), Proceedings of the Seventh Annual Colloquium on Information Retrieval. Philadelphia, Penn.: College of Surgeons. In press.

Simon, Anita and Boyer, E. G. (Eds.), (1968). Mirrors for Behavior: An Anthology of Classroom Observation Instruments. Philadelphia, Penn.: Classroom Interaction Newsletter in cooperation with Research for Better Schools, Inc.

Simon, Anita and Boyer, E. G., (1970). Mirrors for Behavior II: An Anthology of Observation Instruments, Vols. A and B. Philadelphia, Penn.: Classroom Interaction Newsletter in cooperation with Research for Better Schools, Inc.

Smith, B. Othanel, et al, (1967). A Study of the Strategies of Thinking. Urbana, Illinois: Bureau of Educational Research, University of Illinois, Urbana.

Smith, L. M. and Geoffrey, W., (1968). The Complexities of an Urban Classroom: An Analysis Toward a General Theory of Teaching. New York: Holt, Rinehart and Winston, Inc.

Soskins, Vera and Johns, W. F., (1963). The study of spontaneous talk. In Barker, R. G. (Ed.), The Stream of Behavior. New York: Appleton-Century-Crofts.